Environmental Health Perspective
Vol. 18, pp. 97-104, 1976

Assessment of Current Test Procedures

by A. K. Palmer*

The belief that screening tests for teratogenicity are of low predictive value has many supporters who point to the inconsistency with which malformations are induced. However, to fault the test systems when such inconsistency is predictable from both the inherently unstable nature of a malformation and from fundamental principles of teratology, is unrealistic, and, as is shown, perhaps the greater faults lie with the critics.

It is suggested by examples that, if attention was concentrated not on the inconsistent malformations but on more consistent embryopathic effects which in one form or another are always associated with malformations, the predictive value of the screening tests would appear in a more favorable light. Thus, even if malformations are not demonstrated, the range of conditions (dosages) in which they might occur can be determined. Such information, used in conjunction with that obtained from other preclinical studies, can then form a reasonably sound basis for extrapolation to man.

The title of this session suggests this paper should discuss the detection of teratogens, but there seems little profit in this, as the subject has been discussed by everyone who is anyone, every year since 1960.

Should I simply agree with the general opinion that teratogenic tests are of a low predictive value for the conclusive demonstration of teratogenicity? After all, from the fundamental principle that a malformation represents an unstable condition balanced on a knife edge between relatively normal life and death, it is obvious that any number of small factors can tip the balance one way or another. Consequently, even given the mythical, ideal model system, or allowed unrestricted experimentation with pregnant women, looking for teratogens would be worse than looking for a very small needle in a very large haystack.

Then again, what would it mean if we found a foolproof method of detecting teratogenicity? Would it just confirm Karnofsky's law, that any compound can be teratogenic if given to the right species at the right dosage at the right time? Certainly no one seems to worry that we continue to eat salt, despite the fact that it can cause embryotoxicity and teratogenicity in mice; pregnan women still go to the beach, although sand grain cause malformations in chick embryos; and we al breathe oxygen, though it can cause cataracts i premature babies and perhaps may also be im plicated in increased incidences of patent ductu arteriosus in some localities.

At increasingly subtle levels, there exists host of teratogens—illogically referred to as fals positives—which are used in man because th situations in which teratogenicity occurs are ex tremely unlikely to arise in practice. For exam ple, because cortisone causes cleft palate only i mice, corticosteriods are considered fals positives, yet the more powerful fluorinated cor ticosteriods will cause the same kinds of malfor mation in a wide variety of species, includin primates (*1*). Trypan blue is called a false positiv because in rodent species it acts on the yolk sa not present in man; however, as shown by Beck (*2*), trypan blue also causes malformations i ferrets which, like man, do not have a yolk sa placenta. Who would dare to say what would hap pen if we tested these compounds in man to th same degree that we do in animals?

Thus our real question is not whether material is teratogenic but whether it will con stitute a teratogenic hazard when introduced int the human environment. To answer this second

*Department of Reproductive Toxicology, Huntingdon Research Center, Huntingdon TE18 6ES, England.

more pertinent question requires not one, but several investigations, to determine possible beneficial or adverse reactions that may occur. After this exercise, the combined results are assessed to determine whether or not the circumstances required for the beneficial effects coincide with the circumstance in which adverse effects occur.

This second stage in the progess can be likened to building up a jigsaw puzzle (to create a profile of biological activity.) In some cases, a few pieces may give a clear indication of the overall picture; in others, one may have to wait until the last piece falls into place. In this analogy, the teratogenic tests are but one or two small pieces fitting into a section of reporductive toxicology which in turn must be fitted into the larger framework before their significance can be determined.

As an example of the technique, consider the frequently repeated statement that thalidomide would not be detected by current methods. This is a narrow view, based on the negative or doubtful indications of teratogenicity in rats and mice, as on building up the total picture, the low pregnancy rate, low litter size, and poor viability of the few litters obtained in the fertility study (Table 1) would have provided a jarring note against the negative teratogenic and conventional toxicity studies.

Another shock would have arisen from the occurrence of embryotoxicity and a high malformation rate in the rabbit teratology study. The warning given by these awkward pieces of the jigsaw would (or should) have called for further investigation which in these days would quite likely include a primate study leaving no doubt of thalidomide's unique potential (*3*). Thus, it is evident that, if used correctly, the current tests can indicate the hazard of a teratogen even if, initially, they cannot predict its teratogenicity.

With this in mind, I would now like to examine the section of our jigsaw required for testing new drugs to determine their potential effects on development (*4*).

The current system employed by most countries is the three-segment design which for the FDA consists of (I) studies of fertility and general reproductive performance: (both sexes treated continuously from premating through to the end of lactation, one species); (II) studies of teratogenicity and embryotoxicity (pregnant females treated during embryonic organogenesis, two species); (III) peri- and postnatal studies (pregnant females treated during late fetal development through parturition to the end of lactation, one species).

Table 1. Effect of thalidomide in the two-litter test.

Observation	Treatment	Mating					
		Experiment 1			Experiment 2		
		1st	2nd	3rd	1st	2nd	3rd
Conception rate, %	Control	79	74	83	50	62	75
	Thalidomide (200 mg/kg/day)	32	17	6	5	0	0
Litter size at birth	Control	8.3	8.9	8.8	9.2	9.5	9.8
	Thalidomide (200 mg/kg/day)	3.5	4.7	3.0	4.0	0	0

Segment 1: General Fertility and Reproductive Performance

The most widely known general reproductive study, is that of the FDA (Fig. 1). It is conventionally performed in rats, although often it would be equally valid and more economical to use mice, hamsters, or gerbils.

One control and two test groups, each containing 10 males and 20 females, represents a basic minimum. In the U. K., the basic minimum is 12 males, 24 females, and three test groups, which in the long run proves more economical because of the reduction in problems at later stages of assessment. Treatment continues throughout the study after starting 60 days prior to mating for males and 14 days prior to mating for females; a more common variation is to perform two tests, treating the males in one and females in the other. After mating, 10 females per group are killed and examined at mid-pregnancy, for the detection of early effects on implantation and embryonic development. In the UK this sacrifice is delayed to day 20 of pregnancy on the

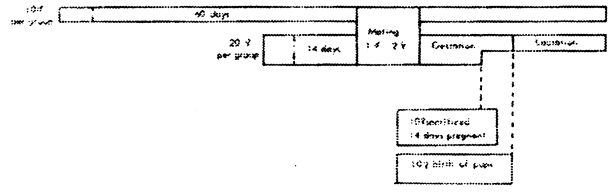

FIGURE 1. Fertility and general reproduction study.

theory that it will allow examination of fetuses for malformations. The remaining females in each group are allowed to bear and rear their young for 21 days to detect effects occurring during late gestation and lactation.

In the UK and to a lesser extent in the USA there have been recommendations that the study be extended by rearing some of the young to maturity in order to detect latent effects on behavior, physiological development, and reproductive capacity. However, practical experience suggests, that introduction of this extension on a routine basis will mostly result in wasted resources, since most compounds that would induce latent effects would already have provided an indication of activity.

At present therefore, extensions would appear to be best confined to instances with *a priori* evidence that the extension would be informative, i.e., for steroids, compounds affecting the neuroendocrine system, addictive drugs, and perhaps certain antibiotics of the types known to induce deafness.

There are a number of design faults in the fertility study which can lead to malpractice. The prolonged treatment period of the male is often incompatible with the concurrent recommendation that compounds be administered by the intended clinical route and that the highest dosage should cause minimal interference with the parental economy. Difficulties particularly arise with unusual routes of administration such as inhalation and instillation into eyes, nose and other body orifices. It is impracticable in the rat to make daily intravenous injections for longer than 4 weeks and impossible with the hamster.

Two other risks associated with prolonged treatment are that subtle cumulative adult toxicity may provoke secondary effects in the later, critical stages of investigation and that the development of detoxification mechanisms and tolerance (which can occur to a marked extent in rats) could mask effects that would arise by *de novo* treatment during fertilization and early embryonic development.

The shorter two-week premating period of females avoids some of the difficulties, but in many cases it is usually too short to demonstrate either tolerance or cumulative toxicity adequately.

The next pitfall in the reproductive study is the interim sacrifice, which together with the fact that only 10 or 12 males are used effectively reduces comparable group size to 10 on most occasions. Thus, given an all-or-none response, such as nonpregnancy, or total litter loss, then, by the laws of probability, statistical significance would be attained only if 40-50% of animals were affected (Table 2).

The importance of this is, that, in practice, the most common primary effect of highly active compounds is the general nonspecific response of nonpregnancy. For example, not only is nonpregnancy the primary effect of the teratogen thalidomide (Table 1), it is also the primary response to haloperidol (Table 3), which affects mating; in turn the effect on mating masks the fact that haloperidol can cause delayed implantation and possibly also affects later physiological development.

Nonpregnancy or, more precisely, pseudopregnancy is the predominant response to guanethidine-like hypotensives, which cause a reversible failure in ejaculation.

Thus, with widely different actions and different hazards, we encounter the same response, i.e., nonpregnancy; therefore, we can anticipate the same response with many compounds. But unlike these highly active compounds, the majority of materials examined in practice are not highly active and, more likely, will give borderline results if any. In these circumstances the necessity to obtain a 40-50% difference to be certain of an effect seems somewhat insensitive.

Table 2. Significance of differences in pregnancy rate with 10 animals per group.

Control success		Test group success Exact test at $p \leq 0.05$			
		One-tailed		Two-tailed	
No.	%	No.	%	No.	%
10	100	6	60	5	50
9	90	4	40	3	30
8	80	3	30	2	20

Table 3. Effect of haloperidol in two-litter test.

Treatment	Conception rate, %	
	Continuous dosing during mating	Withdrawal for a 7 day mating perio
Control	79	79
Haloperidol	22	44
0.1 mg/kg		
1.0 mg/kg	0	39
5.0 mg/kg	0	12

Segment II: Teratology Studies

Teratology studies are by far the best known and most widely performed studies in reproductive toxicology. Basically, requirements call for studies in two species, treatment being applied during the critical phases of organogenesis. This is transposed chronologically to days 6 to 15 of pregnancy in rats and mice and days 6 to 18 of pregnancy in rabbits. (Fig. 2).

In the USA, one control and two test groups each containing 20 rats and mice or 10 rabbits, represents the basic minimum; in the UK and several European countries, three test groups are required. Occasionally, other species may be used, and for larger ones such as pigs, primates, and dogs, smaller group sizes are allowed. The reasons for this are entirely economic and completely unscientific.

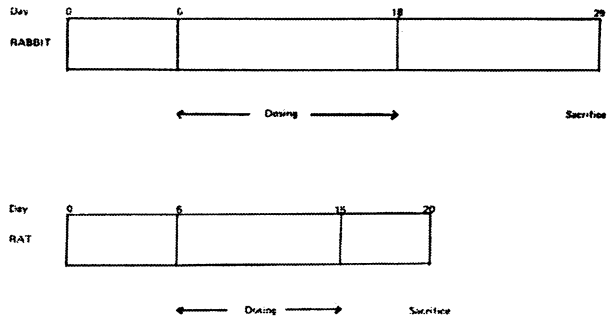

FIGURE 2. FDA segment 2: teratology studies.

Fetuses are usually delivered by hysterectomy one or two days prior to parturition; the numbers of live young and embryonic deaths are recorded. Fetuses are also weighed and examined for external, skeletal, and visceral malformations.

I believe that the design of segment II study is essentially correct for general purposes. That this is sometimes doubted, is mainly due to misuse and misinterpretation (*5*). Why these human errors should occur is due to many reasons, but one of the most common is failure to remember the basic principle that malformations may be induced only under precise conditions, particularly of timing and dosage. (*6*).

Thus, because a relatively few, highly active teratogens are consistently shown by the published literature to provide high rates of obvious malformations, it is believed that all one has to do to determine teratogenicity is to administer a material to a few pregnant animals.

This is not the case, because for new compounds there are no previously published guides to the precise optimum conditions for obtaining malformations; moreover, most compounds that are tested have a low potential for causing malformations. Consequently, the chances of obtaining malformations at a rate high enough to be distinguished against the normal background are extremely remote.

Perhaps, therefore, it would help if we stopped calling these studies teratogenic tests and used the term, "tests for selective embryopathy" or as Wilson puts it, tests for developmental toxicity."

There are several advantages for de-emphasizing the search for teratogenicity. First, it reminds us that we are interested in detecting hazardous materials. In this respect it is just as important to detect embryolethality, retardation of fetal growth, or even increased maternal toxicity (as occurs, for example, with many anti-inflammatory agents or iron dextrans). In other words, we are looking for any effects that would reduce the safety margin predicted by other toxicity tests.

Another advantage of de-emphasizing teratogenicity is that more consistant, objectively determined parameters, such as maternal and fetal weight or numbers of live and dead young, are used to determine whether activity is present. Moreover, in approaching the test in this manner we *do not* reduce the risk of failing to detect hazardous teratogenic compounds. The reason for this is that the unstable, inconsistent nature of a malformation also ensures that other effects occur. For example, most classic teratogens—including the unique thalidomide—will cause embryonic death at the same dosages as those causing malformations (Figs. 3 and 4). Also, all species show a number of variations and anomalies which, because they are less detrimental to survival, occur in more measurable numbers than major malformations. Thus changes in their incidence due to teratogenic activity can be more readily analyzed. The effect of aspirin on the incidence of extra ribs in the rat provides a classic example with its significant dosage-related trend (Table 4). Interestingly, changes in the incidence of these variations are commonly induced at dosages below those causing frank malformations, possibly because some of them are the results of compensatory mechanisms invoked to halt the progress to malformation and death.

Even in the extremely rare instance where embryolethality is not encountered, other clues do exist; for example, with one unique rat teratogen there was a reduction in fetal weight (Table 5).

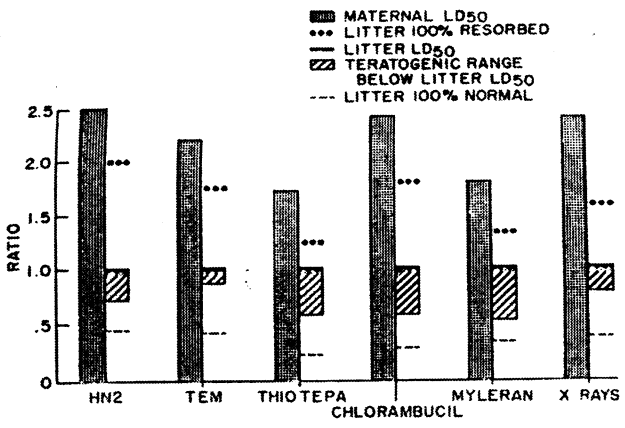

FIGURE 3. The narrow range of dosages for teratogenic effects.

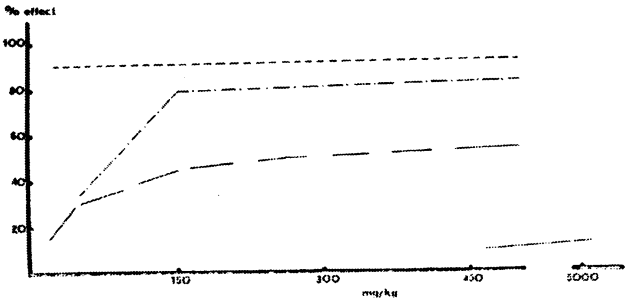

FIGURE 4. Thalidomide dose response: (— —) maternal toxicity; (- -) therapeutic; (- -) resorption; (· - ·) abnormal.

Table 4. Effect of aspirin on the incidence of extra ribs in the CD rat.

Dosage, mg/kg-day	Pups with extra ribs, %
0	10
150	58 *
300	98 *

* $p>0.001$

The unique nature of this teratogen was the absence of the embryolethality shown by classic teratogens (Fig. 3).

On a cautionary note, I must point out that the reliability of these other indicators of embryopathy stem from the use of repeated dosing and several dosages in screening tests, and they may not be so readily induced, nor so necessary, in the single-dose studies of classic teratology.

In using the tests to determine selective embryopathy I can think of no better principles to follow than those outlined by WHO and the Canadian government (*7, 8*).

Two of the main factors in these recommendations are that materials should be administered by the intended clinical route (where possible) and that dosages should be set as follows: (1) the highest dosage should cause a minimal interference with maternal economy (i.e., minimal toxicity); (2) the lowest dosages should preferably cause a clinical (pharmacological) response in the test species; and (3) one or two intermediate dosages should be logarithmically situated between higher and lower dosages. Obviously, it will not always be possible to follow these guidelines to the letter, but with common sense the intent can be followed.

A clinical effect may be difficult to show, so it is best to come down in logarithmic sequence from the high dosage. If the test material has a short half-life, it may be advisable to give several doses in one day; with a depot preparation perhaps at intervals of one or more days. Where cumulative toxicity or rapid development of tolerance occurs, one may have to dose for a shorter period and double up the study to cover the whole susceptible period of organogenesis.

If neither clinical nor toxic effects can be obtained by routes allowing administration of very large doses, one should question whether the material is active or whether the correct species is being used. The term minimal toxic effect should be interpreted broadly and should include extended pharmacological activity. Do not set dosages for repeated dose studies as arbitrary multiples of the single dose LD_{50} value; the results can be fatal, literally as well as figuratively. For example, in an ESSDT cooperative study (*8*), several laboratories tested Myleran, using repeated daily dosages of 1/3, 1/9, and 1/27 of the single dose LD_{50}. Given the nature of the test material it is hardly surprising that most animals at the highest two dosages died and that there were many instances of total litter loss at the lower dosages (*9*).

Having set up the study as a test for selective embryopathy, it is important to assess the results from the same direction, and one of the ways to do this is to examine the pattern and type of dose response. One relatively rare type is the occurrence only of maternal toxicity, usually death. This contains no real risk of teratogenicity but it is advisable to compare with other studies to determine whether pregnancy has altered the level of adult toxicity. The next type of response occurs very frequently, and in this pattern there is increased embryotoxicity but no teratogenicity. More often than not, the em-

bryotoxicity is a secondary consequency of maternal toxicity which can often be indicated by distinct "all-or-none" litter effects. When the embryonic response is close to the maternal response there is no alteration in the predicted safety margin but, if there are marked differences between embryotoxic and maternal toxic dosages, you have either a dangerous material or a marketable abortifacient.

Usually more than one test is required to distinguish this response from the next most common one, which occurs with teratogens. This example (Fig. 5) refers to aspirin but could apply equally to many others, especially the classic teratogens. With this pattern a teratogenic zone occurs just below, or overlapping the dosages causing maternal toxicity. Naturally, with the first dosages chosen, the optimal dosages for teratogenicity may not be encountered, but, because of the associations between malformations and other effects, one knows the range to be explored with a second study. With this type of response one would refer to other studies such as pharmacology or pharmacokinetics to ensure that there was sufficient margin between the clinical dosage and the lowest no-effect dosage in the teratology study. Alternatively, as with anticancer agents, one would need to be assured that the benefits of treatment would justify approaching a teratogenic zone.

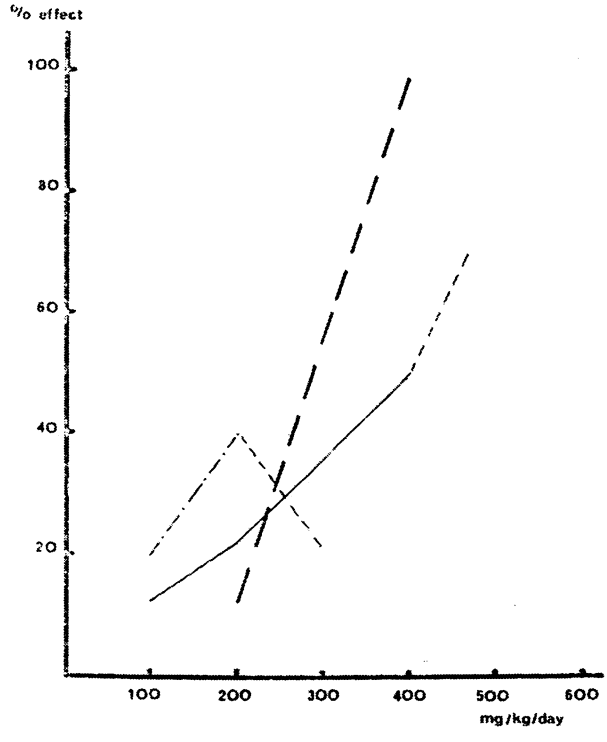

FIGURE 5. Aspirin toxicity: (——) maternal toxicity; (–) resorption; (···) abnormal.

Finally, the dose response of thalidomide (Fig. 4) shows a potential high risk situation. Teratogenicity and embryotoxicity have occurred at much lower dosages than maternal toxicity, indicating selective embryopathy and totally altering the safety margin predicted by other studies.

Segment III: Perinatal and Postnatal Study

Completing the three-segment design is the perinatal and postnatal study (Fig. 6). Basically 20 female rats per group are treated during the last quarter of pregnancy and through lactation. Litter parameters such as growth and development of young are examined from birth through lactation to weaning, at which time animals are killed and examined. In the USA a minimum of two test groups is employed, but the use of three test groups and a control as recommended by the UK and Sweden is preferable.

The virtue of this study is the simplicity of design since it confers great flexibility in introducing the small modifications to make it suitable for a particular test material. The study can be transformed readily to a variety of species, including dogs, pigs, and primates. It can also be modified into a cross-fostering study to distinguish between direct and maternally mediated effects. It is easier than the fertility study to extend for investigation of late development effects.

Perhaps, the values of this test are not fully appreciated unless it is designed and performed in conjunction with the fertility and teratology studies. My best example of this concerns the potent rat teratogens mentioned earlier (Table 5). With this agent a high neonatal mortality rate, in a form of perinatal and postnatal study, provided the first indication that something was wrong (Table 6). Cross-fostering studies showed that the effect was on the fetus, not the parent, and the teratology study subsequently revealed a high incidence of heart malformations.

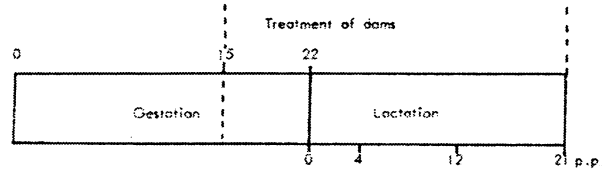

FIGURE 6. Perinatal and postnatal study.

Table 5. Unique teratogenic action in rats.

Group	litters	Mean values per litter: No. of viable young	No. of resorptions	Fetal weight	Visceral malformation, %
1	18	8.1	1.6	3.74	16.7
2	18	10.8 [a]	0.5	3.58 [a]	32.2
3	19	10.1	0.5	3.35 [b]	65.2 [b]
4	20	9.8	1.5	3.28 [c]	88.6 [b]

[a] $p<0.05$, Wilcoxon test.
[b] $p<0.01$.
[c] $p<0.001$.

Table 6. Effect of teratogenic action on postnatal mortality rate: mean values per litter.

Group	At birth: Viable young	At birth: Loss, %	Loss at 21 days postpartum, %
1 (Control)	10.4	8.6	18.2
2	9.9	19.3	40.8
3	5.1 [a]	51.8 [a]	98.5 [a]
4	3.4 [b]	71.9 [b]	100.0 [c]

[a] $p<0.01$, Wilcoxon test.
[b] $p<0.001$.
[c] $p<0.05$.

Conclusion

Having completed my description of the three-segment design for reproductive toxicology it may seem that I have strayed from the expected path. Much of my waywardness is deliberate. Many of the loose ends can only be tied up by reference to other toxicological, pharmacological, and pharmacokinetic studies, which reminds us that for proper extrapolation, results must be placed in the larger framework of biological activity.

I have commented more on the way people perform and assess studies than on the animals or techniques. Again this is intentional, because I believe that many of the so called "faults" of the test are in fact the faults of the investigators and assessors; as just one example from many, Table 7 illustrates how incorrect analysis can make "significant" differences out of nothing.

Thus the only major design change I would make at the present time would be to simplify the general reproductive study by making it consist of equal numbers of males and females per group and omitting the interim sacrifice. Also it could

Table 7. Effect of statistical analysis based on incorrect sample units.

Treatment	Mean fetal weight, g: Male	Female
Control	5.2	5.4
Low dose	5.2	5.4
High dose	5.3*	5.5*
Laboratory standard		
Control range	5.3 4.8-5.9	5.6 4.9-6.0

* $p<0.01$ (Anovar) using fetal units instead of letter units.

be extended to a second generation to make it suitable for pesticides and food additives.

Extension to a second generation may also make it suitable for detecting late developmental effects, as for general purposes at an initial screening stage, I can think of no better guide to whether an animal is functioning correctly than that it will grow and reporduce. Improvement in the predictive value of reproductive toxicity studies will come, not from the introduction of sweeping changes to the three segment design, but from the correct application of minor modifications applicable to the individual test compound, and from a greater understanding of the role of initial screening tests. We should also revert back to following the intent of the guide lines rather than the letter.

No one has found the pot of gold at the end of the rainbow nor, will you find an ideal test system to provide all the answers in one go. Therefore the screening tests should not be considered as an end point but as a starting point. In some cases the results together with those of other studies will justify the risk of performing the definitive experiment in man. In other cases, the pieces of our jigsaw may not fit as well, and further investigations will be required before our definitive study in humans.

Thus, rather than complicate initial screening tests, greater use should be made of secondary stage investigations. By this time, the test compound is not so new, there are clues to indicate the best types of investigation, and there is a familiar framework to establish the perspective of results obtained by more sophisticated methods. Because of these factors, the deficiencies of more specialized techniques can be covered and they can be used to more telling effect. For example pig and primate teratology studies, studies of mechanisms of action, special pharmacokinetic studies in the mother and embryo, and, I suspect

many of the suggested behavior studies, are usually more appropriate at this stage than at initial screening stage.

Many sophisticated techniques have been developed by using compounds of known potential. Overzealous transference of these techniques to the routine initial screening stage of new compounds could lead to their inappropriate usage and thereby their devaluation. The approach to initial screening must be different to that for second and third level investigations. Initial screening tests must be simple and robust and provide a broad basis from which more precise and definitive studies can be launched along the lines indicated by the initial results.

REFERENCES

1. Hendrickx, A. G., et al. Teratogenic effects of triamcinolone on the skeletal and lymphoid systems in nonhuman primates Fed. Proc. 34: 1661 (1975).
2. Beck, F. Comparative placental morphology and function. Environ. Health Perspect. 18: 5 (1977).
3. Delahunt, C. S., and Lassen, L. J. Thalidomide syndrome in monkeys. Science 146: 649 (1964).
4. Palmer, A. K. Some thoughts on reproductive studies for safety evaluation. Proc. ESSDT 14: 79 (1972); Exerpta Medica ICS No. 288.
5. Palmer, A. K. Problems associated with the screening of drugs for possible teratogenic activity. In: Experimental Embryology and Teratology, Vol. 1 D. N. M. Woollam and G. Morris, Eds., Elek, London, 1974, pp. 16-33.
6. Wilson, J. G. Embryological considerations in teratology. In: Teratology: Principles and Techniques. J. G. Wilson and J. Warkany, Chicago Press, Chicago, 1964.
7. World Health Organization. Principles for the Testing of Drugs for Teratogenicity. WHO Tech. Rept. Series, No. 364, Geneva, 1967.
8. Anonymous—The Testing of Chemicals for Carcinogenicity, Mutagenicity Teratogenicity. Canadian Ministry of Health and Welfare, Ottawa, 1973.
9. ESSDT. The evaluation of drugs for foetal toxicity and teratogenicity in the rat. Proc. ESSDT 7: 216 (1966).